Press to start

第十四小隊

M084020005廖冠威　M084020024黃偉豪　M084020045沈詩翰

M084020052陳柏辰 M084020056徐淳郁　M094020025劉有耘　M094020063王弘銘


今天有個任務要指派給鬼殺隊
最近有個東西叫「YouTube」似乎蠻奇怪的
需要你們去了解一下

任務的目標就是
1.調查何種影片受歡迎
2.調查影片收入的高低


我會給你們幾個提示去完成這個任務
以下是任務提示

資料集介紹

前處理

EDA

特徵工程

模型建立

人物情報

第一話

資料集介紹

きめつのやいば
鬼滅之刃
YouTube篇

資料集介紹

資料來源

爬台灣資料

資料時間

11/10-12/25

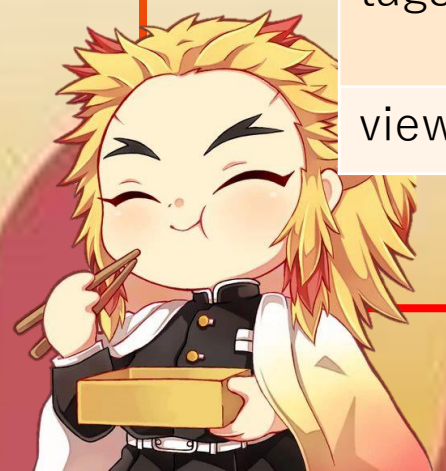
資料筆數

1617筆資料
22個欄位

就讓我先來介紹資料集吧！

資料集介紹

資料欄位-透過Youtube API			
Video_id	影片ID	likes	喜歡數
Trending_date	發燒日期	dislikes	倒讚數
title	影片標題	Comment_count	評論數
Channel_title	頻道標題	Thumbnail_link	影片縮圖連結
Category_id	類別ID	Comments_disabked	是否允許評論
Publish_time	影片發佈時間	Rating_disables	是否允許評分
tags	標籤	Video_erroe_or_removed	影片錯誤或移除
views	觀看數	description	影片描述



第二話

前處理

きめつのやいば
鬼滅之刃
YouTube篇

欄位新增

原欄位	轉換後
tags	Tag_count
description	Desc_length
Trending_date	Trending_day

新增欄位	
income	廣告收入 (views/1000)*cpm
Tag_count	標籤數量
Desc_length	影片描述字數
Trending_day	發燒天數
Like_ratio	喜歡比(likes/dislikes)

將原本欄位轉換成數值資料

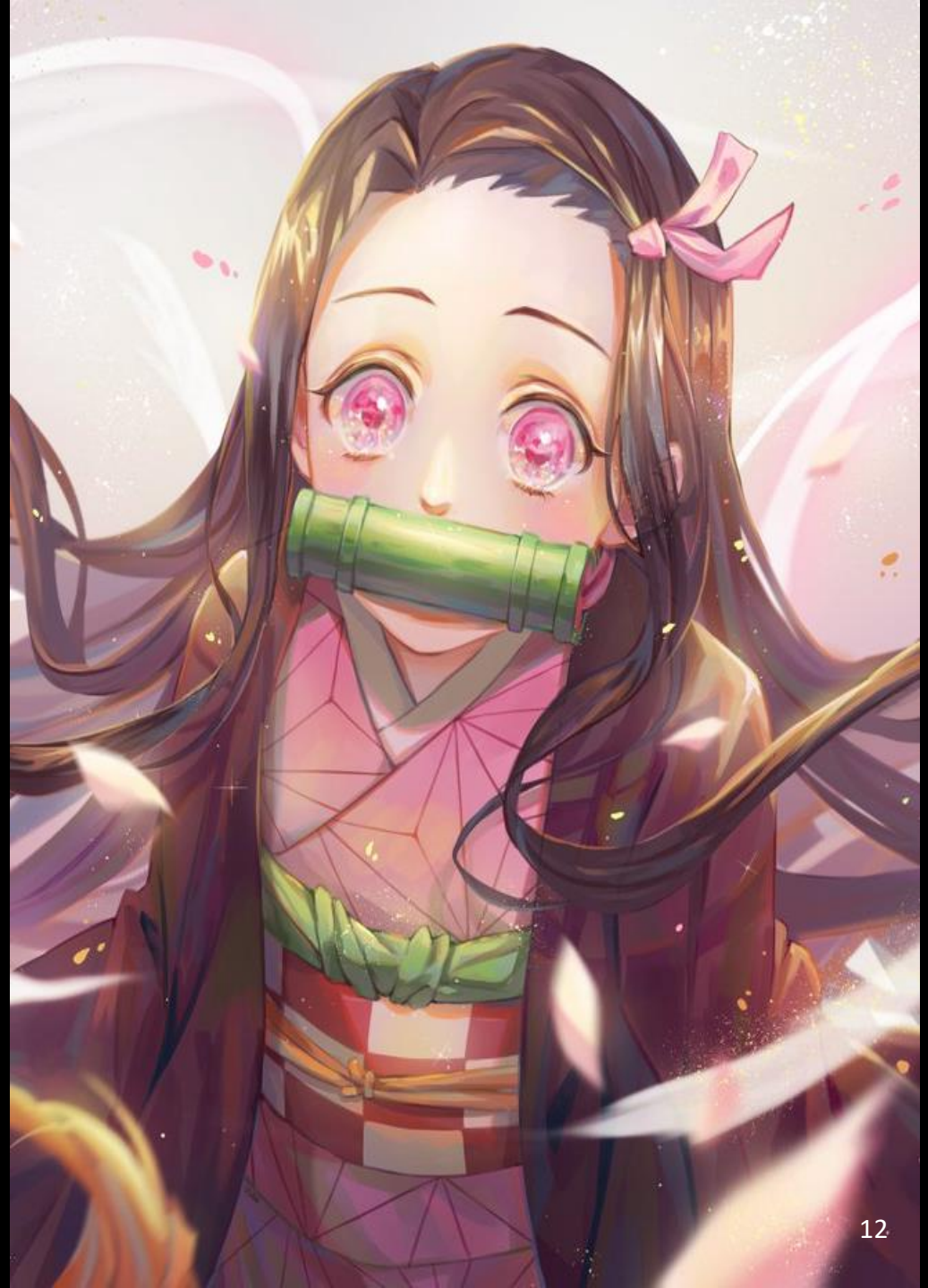
欄位新增

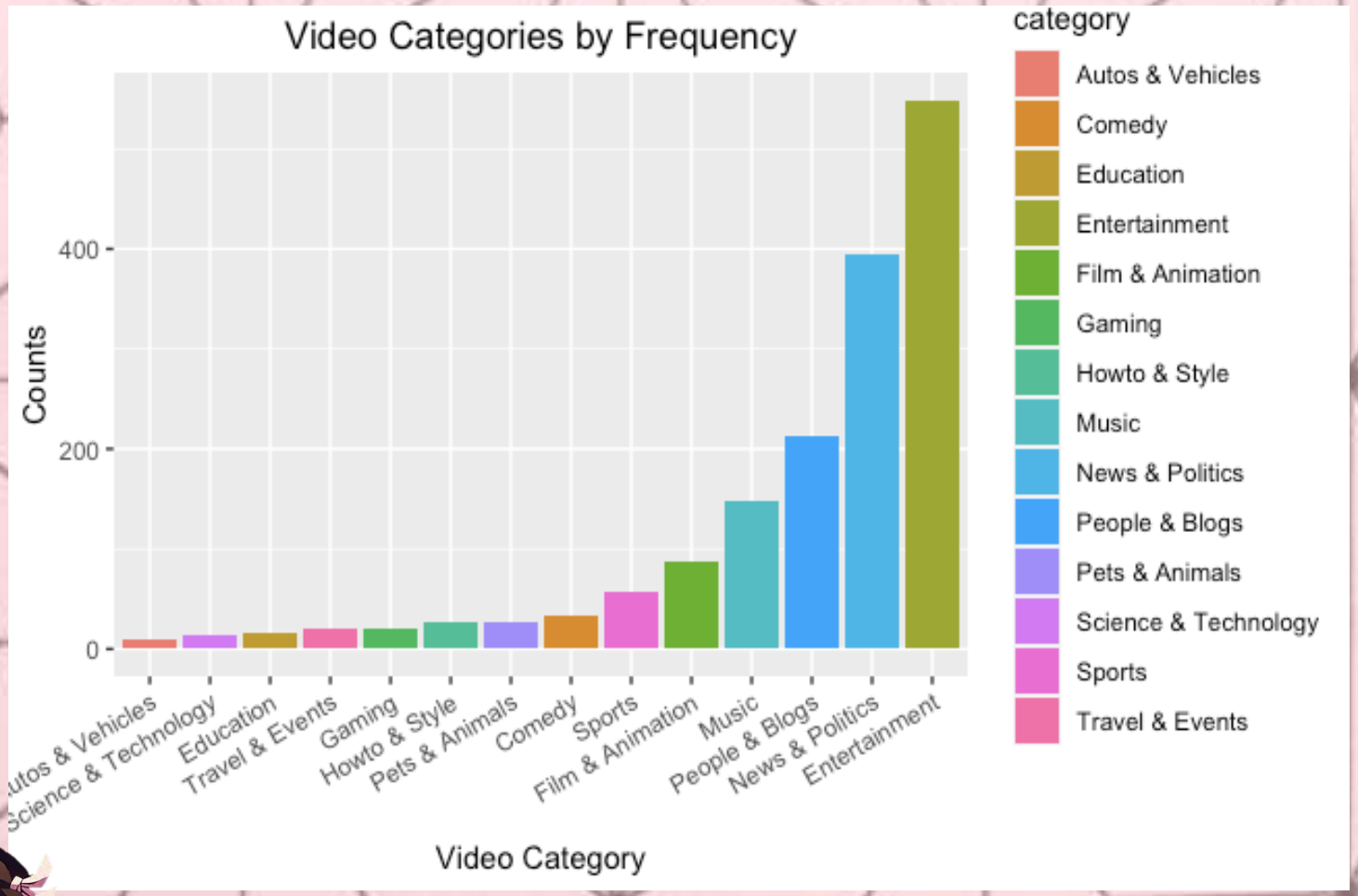
trending_date	tags	description
2020/12/25	大胃王|大胃	喜歡看我吃
2020/12/6	TWICE|I CA	Provided to Y
2020/12/7	BIGHIT|<U+	BTS (<U+BC
2020/12/16	hamtaro gasa	By - Zaty Far
2020/12/12	<U+C5E0><U	[IZ*ONE On
2020/11/20	Sports|SSU|大	本賽事為本
2020/12/2	大政治大爆	#大新聞大爆
2020/12/10	全明星運動	【射箭、排
2020/12/15	這群人|TGO	遠親不如近

income	tag_count	desc_length	trending_day	duration
80.417	74	223	1	1394
1241.50219	2	523	6	206
104918.03	6	1384	17	231
816.89689	37	1519	6	226
1436.90252	59	272	4	269
9.35563	10	756	3	6121
278.86055	67	146	10	3081
870.03937	9	396	16	5792
933.60774	42	1603	14	885
44.09719	10	779	7	762
305.26718	9	122	7	1002

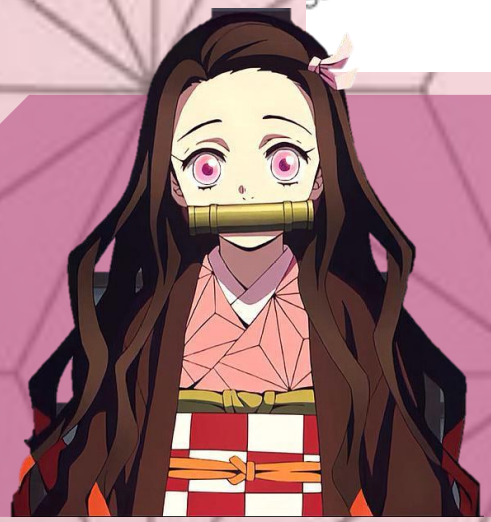
第三話

EDA



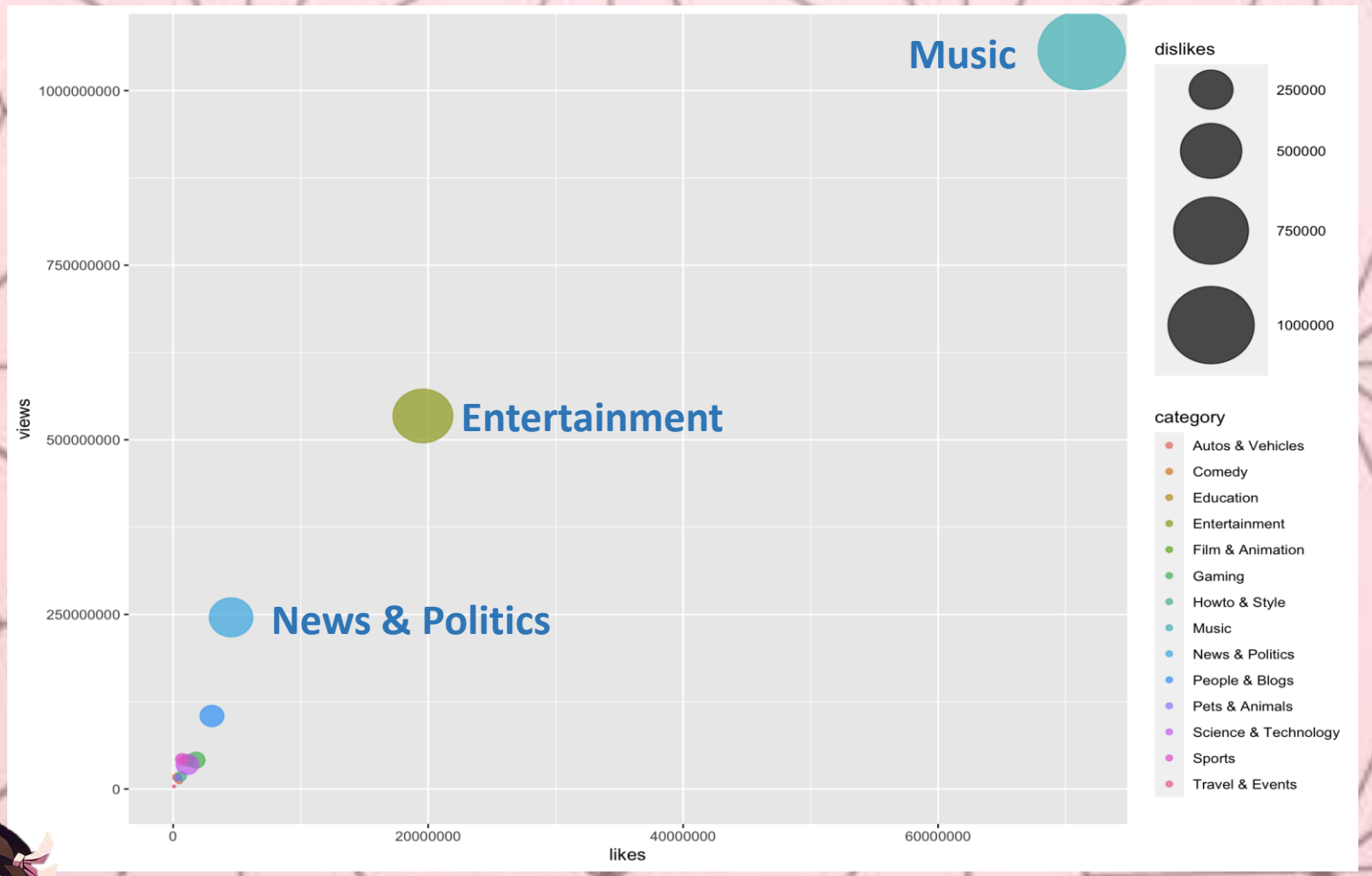


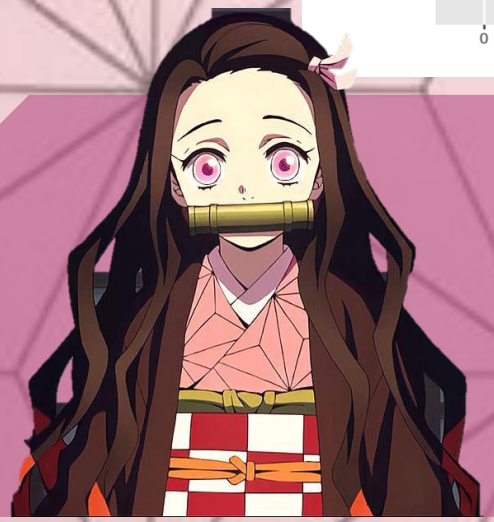
全集中回顧



發燒影片最常見的類型為「娛樂」類別影片

全集中回顧



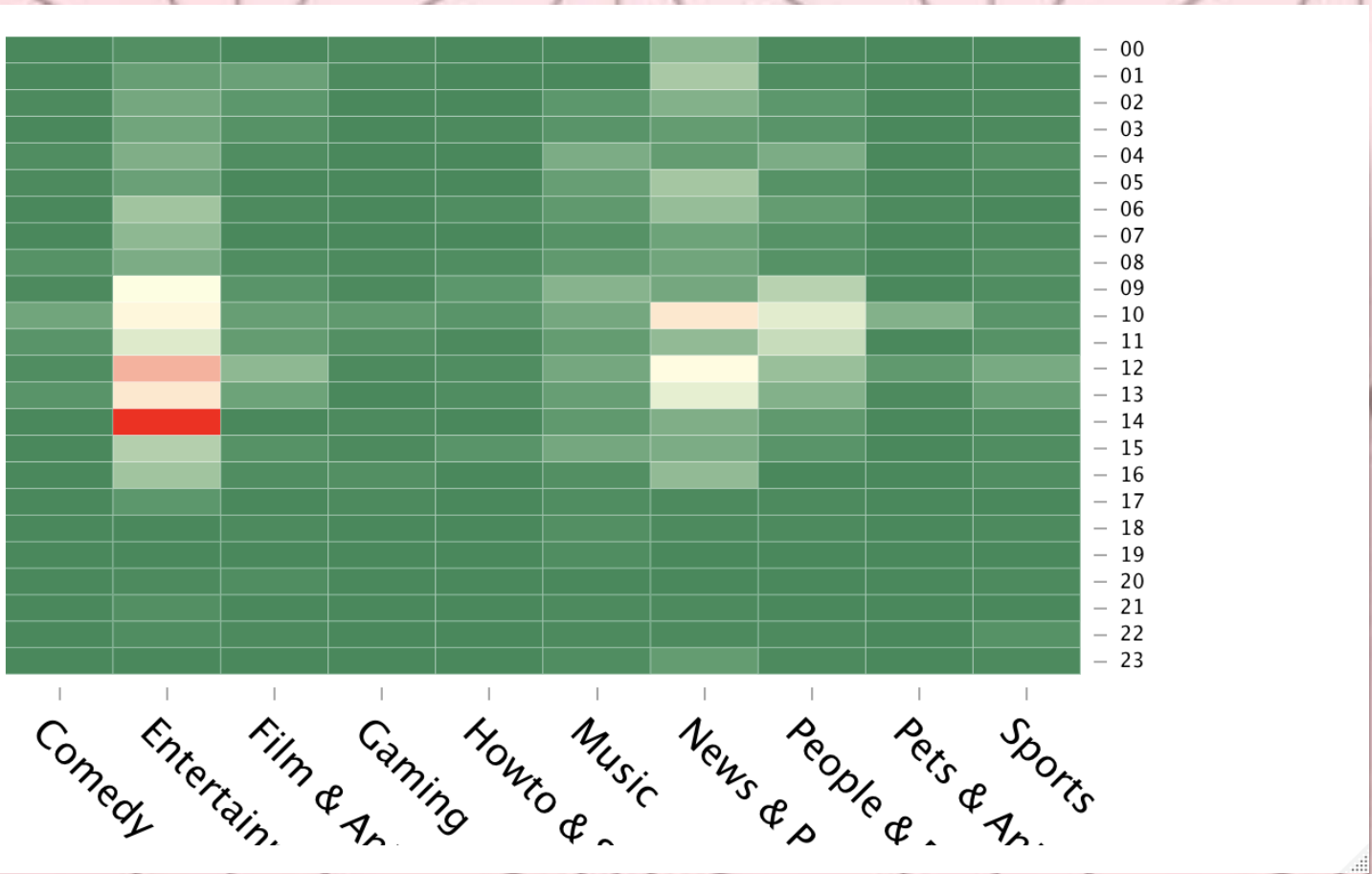


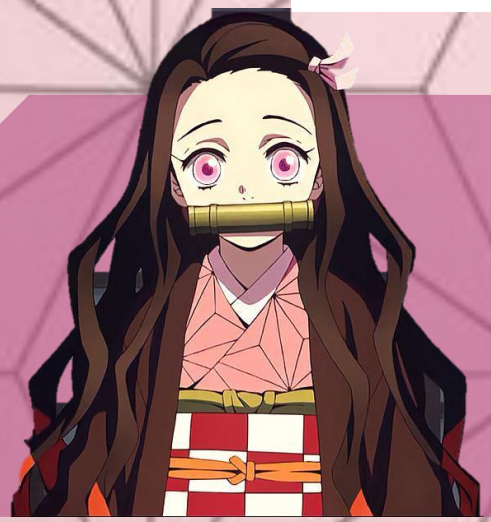
觀看和喜歡數最多的類別
第一為「音樂」其次是「娛樂」

全集中回顧

娛樂類型影片的標籤關鍵字

全集中回顧





影片投放時間
娛樂類別介在9:00-15:00

全集中回顧

根據上述得知，娛樂型影片為現今台灣Youtube主流
因此決定以娛樂型YouTuber為目標

第四話

特徴工程

きめつのやいば
鬼滅之刃
YouTube篇

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins

tw

Filter

category_Autos & Vehicles	category_Comedy	category_Education	category_Entertainment	category_Film & Animation	category_Gaming	category_Howto & Style	category_Music	category_News & Politics
0	0	0	0	1	0	0	0	
0	0	0	0	1	0	0	0	
0	0	0	0	1	0	0	0	
0	0	0	0	1	0	0	0	
0	0	0	0	1	0	0	0	
0	0	0	0	1	0	0	0	
0	0	0	0	1	0	0	0	
0	0	0	0	1	0	0	0	
0	0	0	0	1	0	0	0	
0	0	0	0	1	0	0	0	
0	0	0	0	1	0	0	0	
0	0	0	0	1	0	0	0	
0	0	0	0	1	0	0	0	
0	0	0	0	1	0	0	0	
0	0	0	0	1	0	0	0	
0	0	0	0	1	0	0	0	
0	0	0	0	1	0	0	0	
0	0	0	0	1	0	0	0	
0	0	0	0	1	0	0	0	
0	0	0	0	1	0	0	0	
0	0	0	0	1	0	0	0	
0	0	0	0	1	0	0	0	

資料轉換 — One Hot encoding

將各欄位數據壓縮成0與1的值

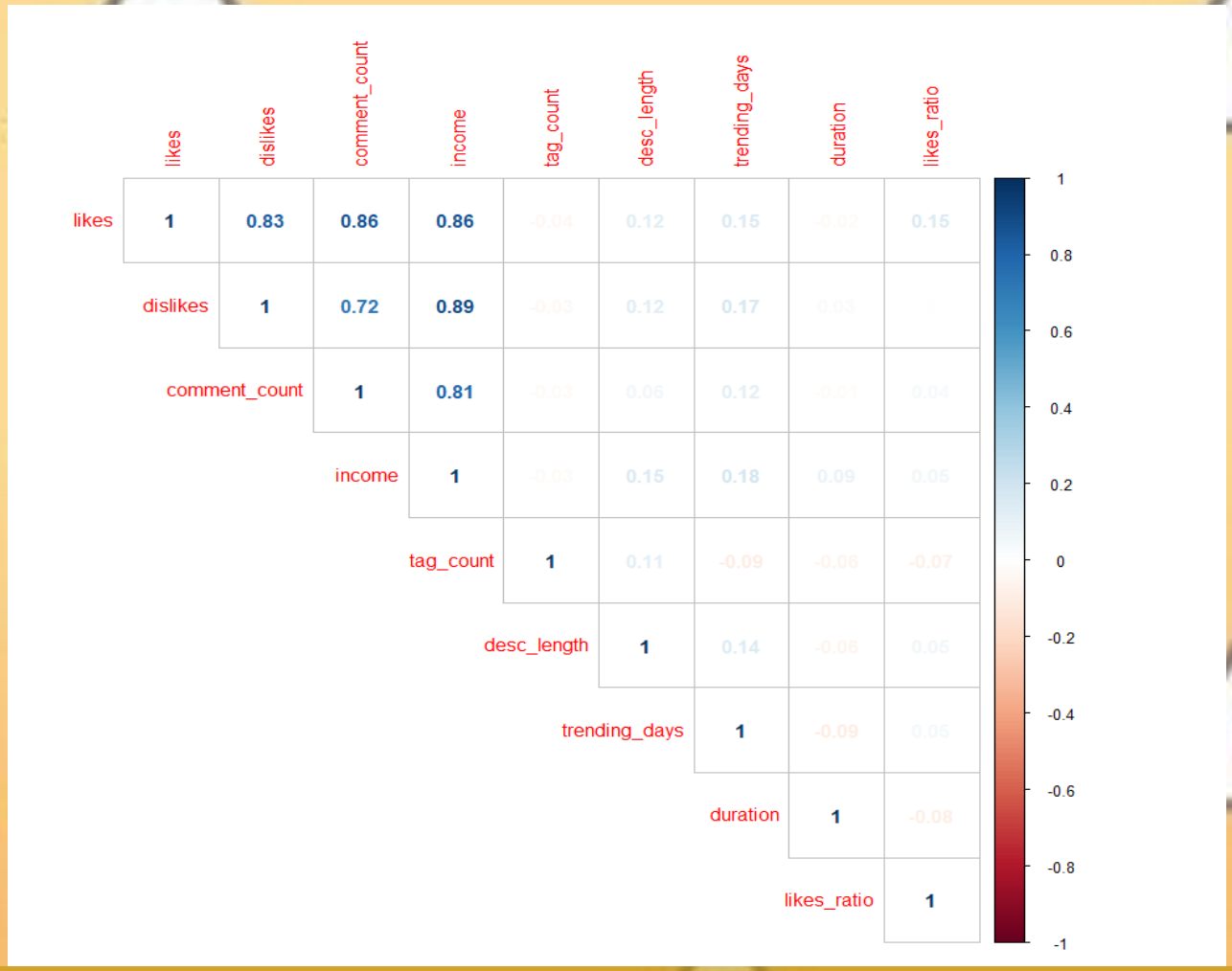
After

video_id	categoryId	likes	dislikes	comment_co	tag_count	desc_length	trending_day	duration	income	likes_ratio
#NAME?	0.77777778	0.00027439	0.00065297	9.95E-05	0.5248227	0.04764957	0	0.0031572	80.417	32.4554455
-4ni0pcKZ4v	0.33333333	0.02313749	0.01843172	0.00443195	0.0141844	0.11175214	0.3125	0.0004432	1241.50219	96.9515959
-5q5mZbe3V	0.33333333	1	1	1	0.04255319	0.2957265	1	0.00050031	104918.03	77.233309
-607OwI4gP	0.33333333	0.00348482	0.00341999	0.00065326	0.26241135	0.32457265	0.3125	0.00048889	816.89689	78.6975425
-6zExiuIotU	0.85185185	0.00895921	0.01049916	0.00132233	0.41843972	0.05811966	0.1875	0.00058712	1436.90252	65.9051724
-8YXCsxF2t	0.59259259	4.44E-06	1.29E-05	1.86E-06	0.07092199	0.16153846	0.125	0.01395609	9.35563	26.5
-9cZ3pt3v28	0.88888889	0.00145676	0.00455136	0.00062211	0.4751773	0.03119658	0.5625	0.00701117	278.86055	24.7201705
#NAME?	0.77777778	0.00092865	0.00380142	0.00083645	0.06382979	0.08461538	0.9375	0.01320449	870.03937	18.8673469
#NAME?	0.85185185	0.00240843	0.00435095	0.00031291	0.29787234	0.34252137	0.8125	0.00199438	933.60774	42.7518574
#NAME?	0.59259259	4.57E-05	5.82E-05	3.72E-05	0.07092199	0.16645299	0.375	0.00171338	44.09719	60.6666667
#NAME?	0.85185185	0.00031382	0.00104733	5.39E-05	0.06382979	0.02606838	0.375	0.00226167	305.26718	23.1419753
#NAME?	0.77777778	0.00079028	0.00223689	0.0002348	0.20567376	0.15641026	0.1875	0.00435428	428.27628	27.2861272
#NAME?	0.77777778	2.94E-05	0.00012284	8.83E-06	0.07092199	0.10320513	0.125	0.00177507	24.57586	18.4736842
#NAME?	0.85185185	0.08769065	0.05909012	0.02024876	0.14184397	0.31901709	0.4375	0.00162657	7950.54736	114.615427
#NAME?	0.92592593	6.79E-05	0.00138998	6.51E-05	0.04964539	0.02692308	0.25	0.00144838	52.89468	3.77209302

正規化—區間壓縮呼吸

將各欄位數據壓縮成[0,1]的值

變數挑選—相關性分析



- **income**
  "likes","dislikes","comment_count","desc_length"
  "trending_days","duration" ,"category"
- **like_ratio**
  "likes","comment_count","desc_length","trending_days","duration"

第五話

模型建立

income

線性迴歸

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     -197.9       71.9  -2.752  0.00601 **
likes          37723.0     3384.6  11.145  < 2e-16 ***
dislikes       52472.2     1742.6  30.111  < 2e-16 ***
comment_count   7189.9      9174.0   0.784  0.43337
desc_length      904.3       293.6   3.080  0.00212 **
trending_days    589.1       201.6   2.923  0.00354 **
duration       11917.3       932.4  12.782  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1322 on 1107 degrees of freedom
Multiple R-squared:  0.7722,    Adjusted R-squared:  0.771
F-statistic: 625.5 on 6 and 1107 DF,  p-value: < 2.2e-16
```

將先前相關性分析挑出的變數進行lm
Train data的MAE為421.4979
Test data的MAE為522.2996

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    65.09      42.06   1.548    0.122
likes       40554.46    2759.96  14.694   <2e-16 ***
dislikes    53588.86    1686.87  31.768   <2e-16 ***
duration    11427.92     931.35  12.270   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1333 on 1110 degrees of freedom
Multiple R-squared:  0.7679,    Adjusted R-squared:  0.7673
F-statistic:  1224 on 3 and 1110 DF,  p-value: < 2.2e-16
```

income

線性迴歸

根據p-value值選擇較顯著的變數進行lm
Mae降低為402.898

線性迴歸

	Overall	IncNodePurity
likes	8.767801	5439774413
dislikes	32.270810	5085354768
comment_count	13.021173	4573992624
tag_count	1.045274	211391607
desc_length	4.993988	396274754
trending_days	3.540197	2391633027
duration	9.436379	981393830
likes_ratio	1.862116	418092068

在進行increased node purity了解變數貢獻度
經過變數挑選後，Mae降為368.0525

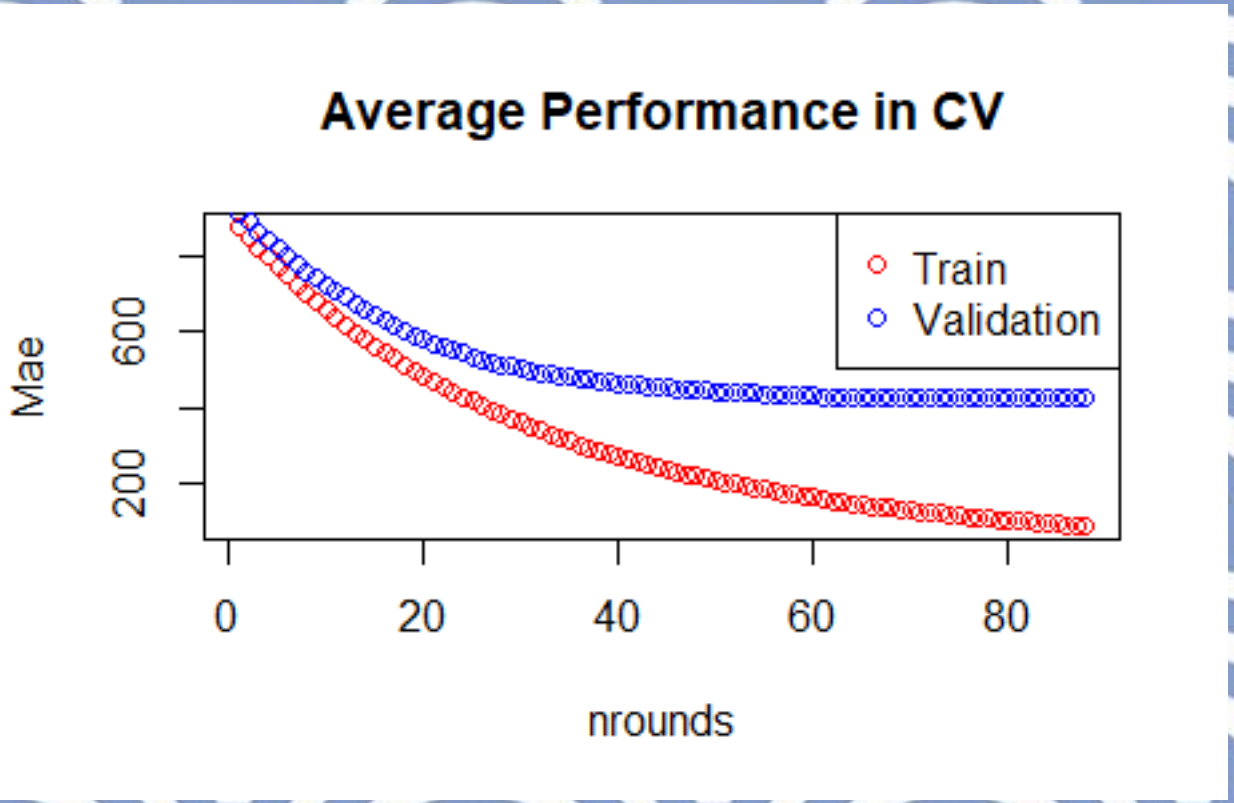
隨機森林

```
> rg_model = ranger(income ~ likes+dislikes+comment_count+desc_length+trending_days+duratio
n , data = train , num.trees = 450 , mtry = 4)
> Mae(train$income , rg_model$predictions)
[1] 327.3853
>
> rg_model2 = ranger(income ~ likes+dislikes+comment_count+trending_days , data = train , n
um.trees = 450 , mtry = 4)
> Mae(train$income , rg_model2$predictions)
[1] 347.6675
```

上面為相關性分析，下面為importance挑選出來的變數

XGBOOST

接著做XGboost，將資料拆為70%訓練集，30%測試集
其中70%訓練集拆分為70%真訓練集，30%真驗證集



接著利用得到的nrounds進行建模
得到的training MAE為131.3754
Validation MAE為493.9305
Testing MAE為216.3893

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     51.336      4.228  12.142  < 2e-16 ***
likes         1438.085    185.576   7.749 2.08e-14 ***
comment_count -1838.976    521.994  -3.523 0.000444 ***
desc_length       5.230     17.279   0.303 0.762180
trending_days    -1.392     11.823  -0.118 0.906284
duration       -147.944     54.097  -2.735 0.006341 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 77.92 on 1108 degrees of freedom
Multiple R-squared:  0.07092,   Adjusted R-squared:  0.06673
F-statistic: 16.92 on 5 and 1108 DF,  p-value: 3.851e-16
```

Likes_ratio

線性迴歸

先利用相關性分析挑選出變數進行lm
Train data的MAE為43.31301
Test data的MAE為38.84577

線性迴歸

```
> new_lm_video_likes = lm(likes_ratio ~ likes +comment_count, data = train)
> MAE(train$likes_ratio,predict(new_lm_video_likes,train))#43.41737
[1] 43.41737
```

再根據p-value值挑選變數，Mae為43.41737

	Overall	IncNodePurity
likes	11.8839347	1964749.0
dislikes	8.0437235	1819494.7
comment_count	7.2336503	1081735.3
tag_count	3.0164335	499612.5
desc_length	1.0220830	585198.4
trending_days	0.6456846	246208.1
duration	2.9569168	896577.0
income	1.8621165	1001190.8

Likes_ratio

線性迴歸

再進行increased node purity了解變數貢獻度
經過變數挑選後進行lm，Mae降為42.20999

隨機森林

```
> rg_model = ranger(likes_ratio ~ likes+comment_count+desc_length+trending_days+duration
 data = train , num.trees = 400 , mtry = 4)
> Mae(train$likes_ratio , rg_model$predictions)#36.52196
[1] 36.52196
> rg_model1 = ranger(likes_ratio ~ likes +dislikes+comment_count+income , data = train ,
m.trees = 400 , mtry = 4)
> Mae(train$likes_ratio , rg_model1$predictions)#7.171722
[1] 7.171722
```

上面為相關性分析
下面為importance挑選出來的變數

等等，以為這樣就結束了嗎？

想完成任務的話，先通過我這一關吧

最終章

上弦之三ー

藝窩座

きめつのやいば

鬼滅之刃

YouTube篇

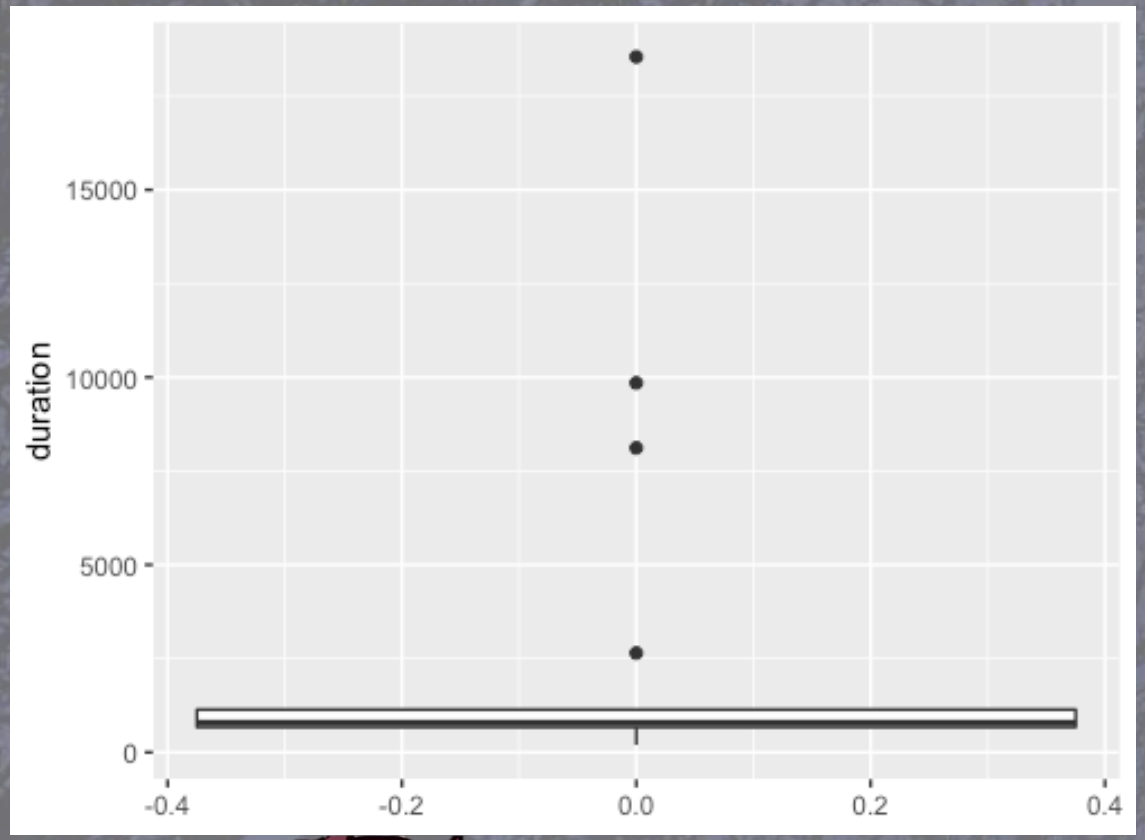
是上弦之三藝窩座噎！！！！
他要我們分析他那沒有破1000訂閱、
觀看時數不足4000小時的頻道
炭治郎現在該怎麼辦？

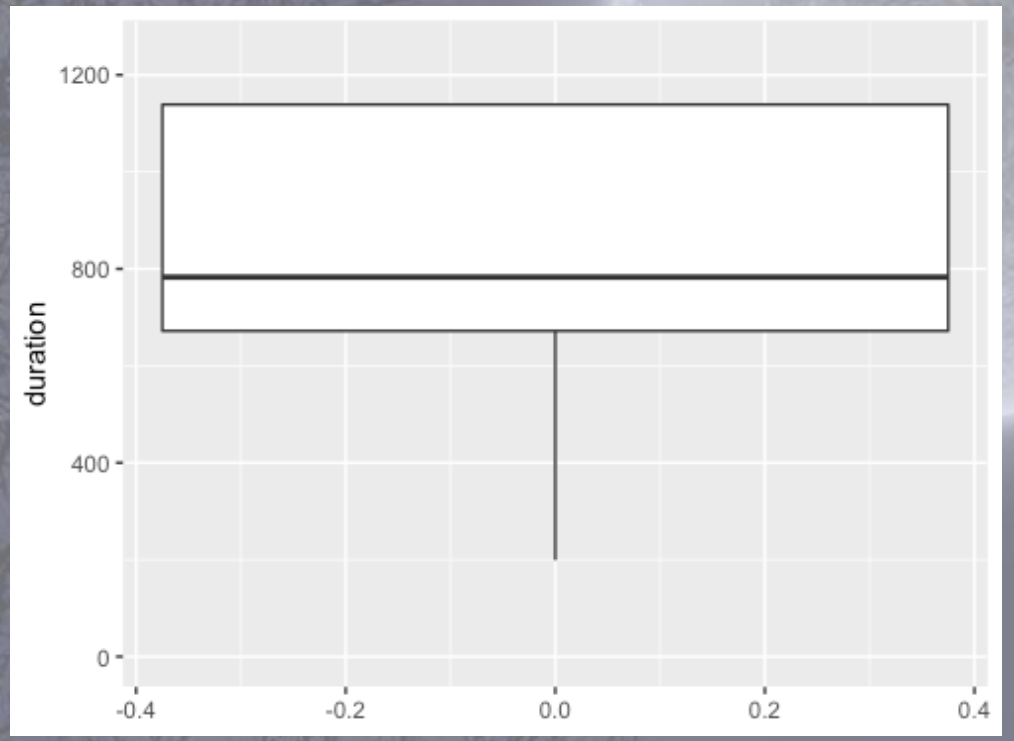
變數挑選

	likes	dislikes	comment_count	income	tag_count	desc_length	trending_days	duration	likes_ratio
likes	1	0.55	0.42	0.43	-0.53	0.4	0.67	-0.52	0.32
dislikes		1	0.86	0.69	-0.34	0.42	0.7	-0.36	-0.41
comment_count			1	0.56	-0.19	0.15	0.46	-0.34	-0.21
income				1	-0.32	0.17	0.63	-0.46	-0.51
tag_count					1	-0.12	-0.41	0.02	0.21
desc_length						1	0.39	-0.44	-0.12
trending_days							1	-0.53	-0.2
duration								1	-0.09
likes_ratio									1

別擔心，利用台灣熱門影片中的「Education」類別
找出影響該類別受歡迎及觀看量的變數供藝窩座參考
並且針對該方向進行改善就好

影片時長





影片時長與觀看量呈高度負相關
將教育類熱門影片時長進行平均
取平均前去除離群值
得出時間為800秒左右

建議：藝窩座將影片時長控制在13分鐘內

文字雲

像這個就是台灣Education類別熱門影片標籤的文字雲分析
發現標籤中皆包含該教育影片內容的主軸

建議：將影片加上R語言或R的tag
或是youtuber的名字直接用本名：藝窩座

影片縮圖

將教育類影片中點擊最高的縮圖取出
縮圖特徵皆包含人像、影片內容概述
建議：藝窩座請不吝展現帥氣的臉蛋
並在縮圖中加入簡短聳動的影片主題

太棒了！這樣任務就算順利結束了吧～

等等，你們忘記最重要的一件事情了

Unit 1 - Introduction to R and RStudio

觀看次數：2,123次 • 2018年5月8日

32 1 分享 儲存 ...

Yihuang Kang
232 位訂閱者

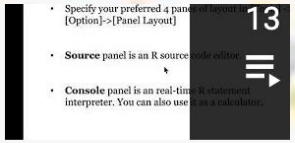
訂閱

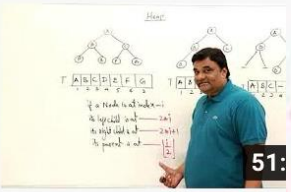
Unit 1 - Introduction to R and RStudio

1 則留言 排序依據

使用以下身分發布公開留言：**Kuan Wei Liao**

Mariuswu Yo 10 個月前
老师课程很好，为什么没有更新了





全
劇
終。